INICIATIVAS DE EVALUACIÓN PARA LA INDIZACIÓN SEMÁNTICA DE LITERATURA MÉDICA EN ESPAÑOL: PLANTL, LILACS, IBECS Y BIOASQ

M. Krallinger, Centro Nacional de Investigaciones Oncológicas (CNIO); Barcelona Supercomputing Center (BSC)

A. Intxaurrondo, Centro Nacional de Investigaciones Oncológicas (CNIO); Barcelona Supercomputing Center (BSC)

E. Primo-Peña, Biblioteca Nacional de ciencias de la Salud (BNCS). Instituto de Salud Carlos III

C. Bojo Canales. Biblioteca Nacional de ciencias de la Salud (BNCS). Instituto de Salud Carlos III

A. Nentidis, National Center for Scientific Research Demokritos, Greece

G. Paliouras, National Center for Scientific Research Demokritos, Greece

M. Villegas, Centro Nacional de Investigaciones Oncológicas (CNIO); Barcelona Supercomputing Center (BSC)

Resumen

El proyecto *Faro de Sanidad* del Plan de Impulso de las Tecnologías del Lenguaje (PlanTL) pretende fomentar el desarrollo de sistemas de procesamiento del lenguaje natural (PLN), minería de textos y traducción automática para español y lenguas cooficiales. Una actividad importante del PlanTL es la organización de campañas de evaluación de sistemas de PLN y minería de textos, un mecanismo que no sólo es clave para evaluar la calidad de los resultados obtenidos por sistemas y algoritmos predictivos, sino que representa un motor fundamental para fomentar el desarrollo de herramientas y recursos de tecnologías del lenguaje.

Debido a la importancia de la literatura para la toma de decisiones en medicina y el volumen considerable de publicaciones en español, el Plan TL, en colaboración con el BSC, el CNIO, la BNCS y la iniciativa BioASQ ha lanzado una tarea competitiva relacionada con la indización automática de la literatura médica en español con términos DeCS. Su fin es generar recursos de etiquetado semántico que sirvan de ayuda a la indización manual. La tarea BioASQ (bioasq.org) de indización semántica biomédica en español se realizará usando resúmenes de artículos de revistas contenidas en las bases de datos LILACS (Literatura Lationamericana en Ciencias de la Salud) y IBECS[1] (Índice Bibliográfico Español en Ciencias de la Salud) como

[1] http://ibecs.isciii.es/

conjunto básico etiquetado y, a partir de ellos, desarrollar los algoritmos de indización automática, facilitando así el desarrollo de modelos de inteligencia artificial.

La evaluación de los sistemas se realiza con la plataforma de BioASQ, mediante un sistema de evaluación continua. En él, se solicita a los participantes que asignen automáticamente términos DeCS a los registros nuevos añadidos a las bases de datos a medida que se hacen públicos, y antes de que se haya completado la indización manual. El rendimiento de indización se calcula comparando indización automática y manual.

Gracias a los resultados de ediciones previas de BioASQ para la indización de PubMed, se ha mejorado este proceso en dicho recurso. Esta tarea de indización biomédica en español servirá para generar recursos comparables para indizar LILACS e IBECS y otros conjuntos documentales.

Palabras clave**:** indización automática de textos; anotación semántica; minería de textos

Abstract

The health flagship project of the Plan for the Advancement of Language Technology (PlanTL) tries to promote the development of natural language processing systems (NLP), text mining and machine translation resources for Spanish and co-official languages. There is a growing demand for a better exploitation of datasets generated by clinicians, especially electronic health records, as well as the integration and management of this kind of data in personalized medicine platforms integrating also information extracted from the literature. In this context, the PlanTL collaborates in the organization of evaluation efforts of clinical NLP and text mining systems, a key mechanism to evaluate the quality of results obtained by such automated systems and a fundamental mechanism to promote the development of tools and resources related to language technologies.

Given the importance of literature for medical decision-making and the growing volume of Spanish medical publications, the TL Plan, in collaboration with the BSC, CNIO, the Biblioteca Nacional de Ciencias de la Salud and the BioASQ team have launched a shared task on automatic indexing of abstracts in Spanish with DeCS terms. The aim of this tracks is to generate semantic annotation resources that can be used to assist manual indexing. The Spanish biomedical semantic indexing track of BioASQ (bioasq.org) will rely on abstracts of journals contained in the LILACS databases as a basic Gold Standard manually labeled benchmark set for the development of automatic indexing algorithms particularly those based on artificial intelligence language models.

The evaluation of participating systems is done through the BioASQ platform, which requests results in a continuous evaluation process, i.e. automatically asking for DeCS term assignment for newly added documents to LILACS, as they are made public, and before the manual indexing results are publicly released. The indexing performance in BioASQ is calculated by comparing automatic indexing against manual annotations.

Thanks to the results of previous editions of BioASQ for indexing PubMed, the MeSH indexing process of this resource was considerably improved. This novel effort on medical indexing in Spanish will serve to generate comparable resources to semantically index not only LILACS but also other health databases and repositories in Spanish.

Key words: semantic annotation; text mining; automatic indexing

Introducción

Debido al creciente volumen de publicaciones biomédicas y médicas, junto con la disponibilidad de bases de datos bibliográficas centralizadas de fácil acceso para sistemas de tecnologías del lenguaje, como es el caso de PubMed, el campo de la minería de textos y procesamiento del lenguaje natural aplicado al dominio biomédico ha experimentado una evolución rápida y productiva, resultando en todo tipo de aplicaciones y recursos software. Este tipo de recursos, casi de forma exclusiva se han desarrollado para procesar textos publicados únicamente en inglés. Los intentos de procesar documentos en otros idiomas ha atraído mucha menos atención a pesar de su evidente interés práctico, en especial para artículos comprendidos dentro de disciplinas mas cercanas al ámbito clínico. Sin embargo, el considerable número de publicaciones médicas escritas en español, genera una necesidad apremiante de facilitar un acceso mas eficaz a la información descrita en estos contenidos mediante herramientas de minería de textos y sistemas de recuperación de información mas sofisticados. Cabe destacar que la literatura médica, y las herramientas desarrolladas para su procesamiento también han sido claves para el desarrollo de recursos terminológicos y sistemas de procesamiento de textos clínicos y de historia clínica electrónica (HCE).

Para abordar este asunto, la Secretaría de Estado para el Avance Digital[2] encargó las actuaciones de apoyo técnico especializado para el desarrollo del Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) en el ámbito de la biomedicina.

El proyecto que presentamos se enmarca dentro del Plan TL de la Agenda Digital para España [1], aprobada en febrero de 2013 como la estrategia del Gobierno para

[2] Secretaría de Estado para el Avance Digital del Ministerio de Economía

desarrollar la economía y la sociedad digital. Esta estrategia se configuró como el paraguas de todas las acciones del Gobierno en materia de Telecomunicaciones y de Sociedad de la Información y marca la hoja de ruta en materia de Tecnologías de la Información y las Comunicaciones (TIC) y de Administración Electrónica para el cumplimiento de los objetivos de la Agenda Digital para Europa[3].

Para la puesta en marcha y ejecución de la Agenda se definieron diferentes planes específicos entre los que se encuentra el Plan TL que tiene como objetivo fomentar el desarrollo del procesamiento del lenguaje natural y la traducción automática en lengua española y lenguas co-oficiales. Para ello, el Plan TL define medidas que: (i) aumenten el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas co-oficiales; (ii) impulsen la Industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria; y (iii) incorporen a la Administración como impulsor del sector de procesamiento de lenguaje natural.

Una actividad fundamental del Plan TL es la organización de campañas de evaluación de sistemas de PLN y minería de textos, un mecanismo que no sólo es clave para evaluar la calidad de los resultados obtenidos por sistemas y algoritmos predictivos, sino que representa un motor fundamental para fomentar el desarrollo de herramientas y recursos de tecnologías del lenguaje [2] [3].

Debido a la importancia de la literatura para la toma de decisiones en medicina y el volumen considerable de publicaciones médicas en español, el Plan TL, en colaboración con el BSC[4], CNIO[5], la BNCS[6] y la iniciativa BioASQ[7], ha lanzado una tarea competitiva relacionada con la indización automática de la literatura médica en español con términos del tesauro DeCS[8] (Descriptores en Ciencias de la Salud). El objetivo es fomentar el desarrollo de sistemas eficientes de indización automática que puedan servir de ayuda a la indización manual, y por consiguiente a una recuperación de información con mejoras en términos de cobertura y precisión. La metodología y recursos que se generen para la indización semántica de la literatura médica en español pueden servir como principio para el enriquecimiento semántico de otro tipo de textos, como webs de salud, guías de practica clínica, tesis y publicaciones académicas, publicaciones de sociedades medicas y asociaciones de pacientes o indización de la HCE.

[3] https://ec.europa.eu/
[4] Barcelona Supercomputing center www.bsc.es
[5] Centro Nacional de Investigaciones Oncológicas www.cnio.es
[6] Biblioteca Nacional de Ciencias de la Salud - Instituto de Salud Carlos III http://www.isciii.es/bncs
[7] http://bioasq.org/
[8] http://decs.bvs.br/E/homepagee.htm

La tarea BioASQ (bioasq.org) de indización semántica biomédica en español, al igual que en las campañas anteriores centradas en PubMed, se realizará usando resúmenes de artículos. En este escenario, los resúmenes provienen de la base de datos LILACS[9] (Literatura Lationamericana en Ciencias de la Salud) e IBECS[10] (Índice Bibliográfico Español en Ciencias de la Salud) como conjunto básico etiquetado para el desarrollo de los algoritmos de indización automática [4]. Estos conjuntos de resúmenes etiquetados con el vocabulario controlado de DeCS sirven como datos de entrenamiento y validación de algoritmos basados en aprendizaje de maquina e inteligencia artificial, los cuales intentan generar modelos predictivos que a su vez asignaran términos DeCS candidatos a nuevos resúmenes.

La evaluación de los sistemas que participan en esta tarea competitiva se realiza mediante un conjunto de evaluación, métricas y formatos comunes lo que permite poder comparar de forma transparente los resultados obtenidos. Los resultados de las estrategias que participan en esta tarea se evaluarán sistemáticamente usando la plataforma BioASQ mediante una evaluación continua. En él, se solicita a los sistemas participantes que asignen automáticamente términos DeCS a los documentos nuevos añadidos a las bases de datos a medida que se hacen públicos, y antes de que se haya completado la indización manual. El rendimiento de indización se calcula comparando la indización automática con la manual.

Gracias a los resultados de ediciones previas de BioASQ para la indización de artículos en inglés en PubMed, se ha mejorado este proceso en dicho recurso. Esta tarea de indización médica en español servirá para generar recursos comparables para indizar LILACS, IBECS y otros conjuntos documentales en español.

En adelante el artículo se organiza de la siguiente manera: la primera sección ofrece una rápida revisión de la extracción de información en biomedicina; la segunda sección describe los objetivos y el funcionamiento típicos de las campañas de evaluación; la tercera sección se centra en las campañas de BioASQ celebradas hasta la fecha; la cuarta sección describe la nueva campaña de indización propuesta para los artículos en español y finalmente, en la última sección, se discuten los resultados esperados.

La extracción de información en biomedicina

Debido al gran volumen de documentos digitales disponibles, los mecanismos de recuperación de textos eficaces y eficientes son de suma importancia. Un aspecto crucial de estos mecanismos es la indización semántica: la descripción precisa del contenido de los documentos con términos extraídos de un vocabulario controlado tipo tesauro, típicamente estructurado de forma jerárquica (vocabulario

[9] http://lilacs.bvsalud.org/es/
[10] http://ibecs.isciii.es/

estructurado). Las etiquetas y la taxonomía se pueden usar, por ejemplo, en los motores de búsqueda para recuperar documentos cuyos conceptos corresponden a los términos de la consulta (o sus sinónimos, hipónimo, etc.) u organizar jerárquicamente los documentos recuperados.

Tradicionalmente, esta anotación se realiza de forma manual de acuerdo con el vocabulario de un dominio y siguiendo unas normas de indización previamente establecidas. La National Library of Medicine de los Estados Unidos (NLM), la biblioteca biomédica más grande del mundo, emplea a expertos biomédicos para indizar artículos de revistas biomédicas etiquetándolos con conceptos del tesauro Medical Subject Heading[11] (MeSH). En España, la BNCS, indiza manualmente la base de datos Índice Bibliográfico Español de Ciencias de la Salud[12] (IBECS) utilizando el tesauro (DeCS)[13], traducción del MesH, utilizado para indexar artículos de revistas biomédicas, libros, informes técnicos y otros materiales con el objetivo de mejorar el proceso de búsqueda.

Esta anotación manual implica costos significativos en tiempo y dinero. Además, el constante aumento de texto digitalizado (y su diversidad) hace cada vez más necesario el uso de indización automatizada; la asignación de descriptores por parte de sistemas informáticos en los que se utilizan diferentes algoritmos para determinar qué términos usar en los descriptores. La indización automática puede servir incluso como tecnología de asistencia para los indizadores humanos.

Etiquetar documentos con conceptos puede verse como un problema de clasificación, en el que los conceptos se consideran clases y los documentos instancias a clasificar. Para este tipo de tare se han propuesto algoritmos especializados de aprendizaje automático y medidas de evaluación para la clasificación jerárquica. Por ejemplo para el concepto ‘Infección Hospitalaria’ con el identificador DeCS D003428, la tarea de clasificación consistiría en clasificar de forma binaria si un documento es o no relevante para este concepto.

Las campañas de evaluación como motor de impulso detrás de las TL

Las campañas de evaluación (conocidas como *shared tasks* en inglés) son comunes en PLN y tienen una larga tradición[14]. Una campaña de evaluación generalmente involucra a cuatro agentes diferentes: (1) los organizadores de tareas o campañas, (2) los grupos de expertos que generan los datos relevantes para las tareas, (3) los participantes o sistemas que participan en las tareas y (4) los usuarios finales de los sistemas generados. Los organizadores de las campañas identifican un reto o tarea

[11] https://www.ncbi.nlm.nih.gov/mesh
[12] http://ibecs.isciii.es/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&base=IBECS&lang=e
[13] http://decs.bvs.br/E/homepagee.htm
[14] Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics. 2015 May 1;17(1):132-44.

específica del dominio que es importante en biomedicina o salud, siendo difícil de resolver y que se considera de alto impacto.

Muchas tareas se diseñaron para satisfacer las necesidades del mundo real en la investigación biomédica, por lo que suele ser una práctica importante y común que los organizadores de tareas incluyan a los usuarios finales en esta etapa de planificación.

Una vez que se determina un reto de dominio, los organizadores de las tareas examinan los recursos existentes para recopilar los materiales apropiados para preparar los datos que se utilizarán durante la fase de desarrollo y la evaluación de los sistemas propuestos.

Generalmente se reclutan expertos en el área temática para anotar (indizar en nuestro caso) manualmente los documentos relevantes que se utilizarán en la campaña. Las anotaciones manualmente generadas se utilizan como datos de referencia (llamados Gold Standards) contra los cuales se comparan los resultados generados por los sistemas de los participantes. Normalmente, los indizadores necesitan hacer referencia a fuentes de conocimiento externas o bases de datos cuando producen los Gold Standards. En caso de las campañas de indización se utiliza MeSH para el inglés.

El Gold Standard generado suele dividirse en dos subconjuntos de datos: uno para el entrenamiento de los sistemas (training) que se desarrollan en la campaña y otro para la evaluación final (test).

Las campañas suelen anunciarse en convocatoria abierta con el fin de facilitar la participación tanto de grupos académicos como de empresas. Los participantes normalmente tienen unos meses para implementar sus sistemas preliminares basados en los datos de entrenamiento distribuidos por los organizadores y anotados por los expertos del área temática.

Al final, los participantes tienen unos días para presentar los resultados de las pruebas y los sistemas participantes se evalúan utilizando métricas de evaluación específicas de la tarea (por ejemplo, precisión, recuperación, medida F, clasificación recíproca) en comparación con el humano.

Las primeras campañas se iniciaron en los Estados Unidos por el NIST[15] en colaboración con DARPA[16] en 1987 y se centraron en el procesamiento del habla [5] Desde entonces, las campañas de evaluación se han convertido en una forma exitosa de impulsar la investigación al tiempo que constituyen un paso importante

[15] National Institute of Standards and Technology https://www.nist.gov/
[16] Defense Advanced Research Projects Agency https://www.darpa.mil/

hacia la estandarización en aspectos claves como son los formatos compartidos, las métricas utilizadas y los criterios de evaluación. Los organizadores de las campañas preparan los conjuntos de datos, definen los criterios de evaluación, clasificar los sistemas, etc., y todo esto ha favorecido el desarrollo de buenas prácticas y estándares *de facto.* De entre los efectos positivos de las campañas de evaluación podemos destacar:

- Evaluaciones objetivas: Todos los sistemas participantes usan los mismos datos, esto facilita la comparación entre sistemas y permite evaluaciones objetivas.
- Generación de nuevos recursos: Las campañas suelen implicar el desarrollo de nuevos recursos que quedan disponibles.
- Reproducibilidad: Cuando los datos quedan disponibles, nuevos investigadores pueden medir el rendimiento de sus sistemas frente a datos de campañas anteriores, lo que permite la reproducibilidad de experimentos.
- Identificación de retos: Las tareas compartidas, especialmente aquellas con un gran número de participantes, ayudan a indicar la necesidad de abordar un problema en particular y a señalar los desafíos relevantes.

BioASQ

BioaASQ es un proyecto que tiene como objetivo promover la investigación en sistemas de información que sean capaces de responder preguntas de contenido biomédico. Para promover la investigación en esta área, BioASQ organiza los llamados *desafíos*, en los que participan algunos de los grupos de investigación más conocidos del mundo [6].

La indización automática de artículos biomédicos es una de las tareas de la campaña anual de BioASQ. El objetivo de esta tarea es identificar los descriptores que propone MeSH que mejor describen un artículo de PubMed.

La National Library of Medicine de los Estados Unidos anunció en 2014 los beneficios significativos que obtuvieron de su participación en las campañas de BioASQ, destacando las mejoras obtenidas en sistema *Medical Text Indexer* con el que la NLM indexa la literatura biomédica basada en el tesauro MeSH [7].

La Figura 1 muestra el éxito de participación en las diferentes campañas anuales de BioASQ previamente realizadas para datos en inglés. Cabe destacar el impacto que ha tenido esta iniciativa para la comunidad de grupos de investigación en sistemas de recuperación en el ámbito de salud, con casi 700 usuarios registrados en el 2017 y 10 equipos participantes en la tarea de indización con 31 sistemas diferentes. La Figura 2da una visión general de la distribución geográfica de los participantes y,

finalmente, la Figura 3 muestra cómo las sucesivas tareas anuales han ayudado a promover y mejorar la investigación y el estado del arte en el ámbito de la indización.

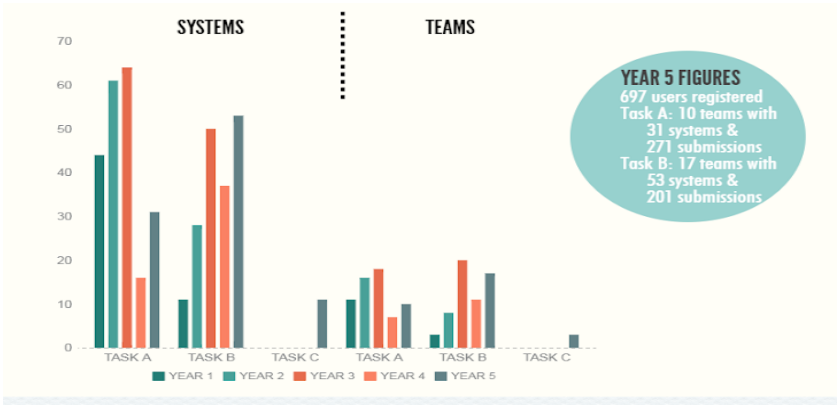

Figura 1 Participación en las diferentes campañas de BioASQ

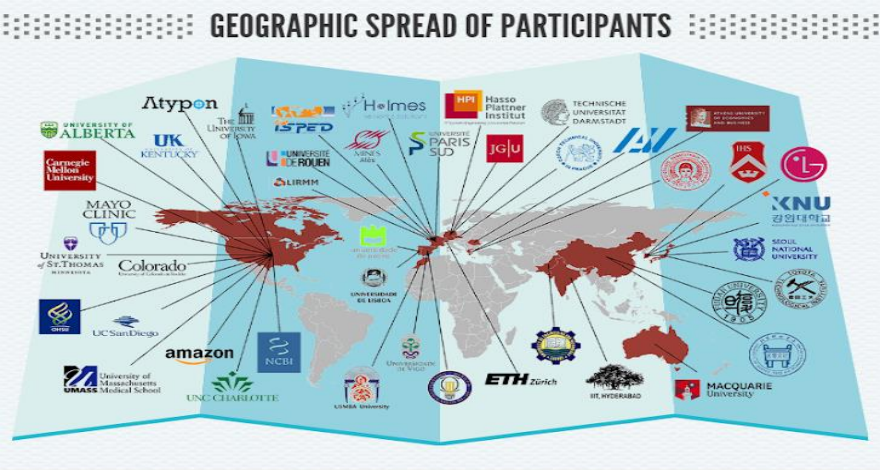

Figura 2 Distribución geográfica de los participantes

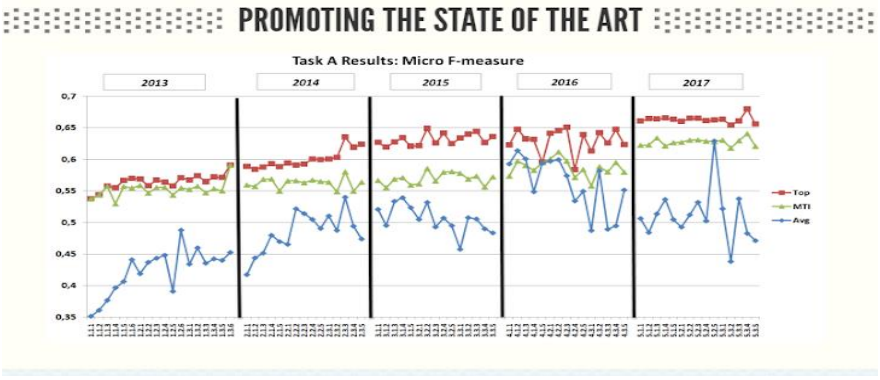

Figura 3 Avances obtenidos

Tarea de indización LILACS e IBECS en BioASQ

Durante las últimas décadas del siglo XX las revistas científico-técnicas editadas en América Latina, España y Portugal se encontraban en una situación que podría calificarse como delicada. Los principales problemas que enfrentaba este grupo de revistas se referían, a grandes rasgos, a la escasa visibilidad, discreto impacto en la producción global del conocimiento (la mayoría de las veces local o institucional), falta de registro, insuficiente indización en los principales índices nacionales, regionales e internacionales, dificultades para la distribución e insuficiente respaldo institucional. En definitiva, muchas de las revistas se encontraban con serias dificultades para sobrevivir y los investigadores tenían problemas para conocer qué, quién, dónde y cómo se investiga y publica en cada área del conocimiento (Bojo Canales C. Las revistas científicas iberoamericanas. Boletin MEDES, 2010. Disponible en: https://www.fundacionlilly.com/global/img/pdf/actividades/medes/boletines_medes/boletin-medes-2010_5.pdf)

Para tratar de superar esta situación diversas instituciones del ámbito iberoamericano pusieron en marcha iniciativas destinadas a registrar y dar visibilidad y difusión a las revistas científicas de la región. Una de las más importantes fue la puesta en marcha de LILACS, Literatura Latinoamericana en Ciencias de la Salud, una base de datos bibliográfica coordinada por BIREME (Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud), que, en términos generales, abarca la mayor parte de literatura científica médica, producida por los países de la región desde 1982 en adelante y que incluye no sólo referencias de artículos de revista sino también monografías, tesis, actas de congresos...etc).

LILACS se desarrolla de forma colectiva regional; BIREME es el centro coordinador que define metodologías de trabajo y forma al personal de los países participantes. Cada uno de los países participantes contribuye a LILACS con su propia literatura científico médica.

LILACS utiliza el tesauro DeCS, traducción del MESH de la NLM de EEUU, para indizar los contenidos recogidos. La indización basada en un tesauro es una de las tareas documentales más complejas y resulta clave para alcanzar un grado aceptable de precisión en la recuperación de información en la base de datos.

La Tabla irecoge algunas cifras sobre el contenido de artículos en español indizados en LILACS (datos extraído de la web de LILACS en junio del 2018).

Número de artículos con resumen en español	284,502
Número de artículos con resumen en español y códigos DeCs asignados	189,693
Número total de códigos usados	1,472,746
Media de códigos DeCs por artículo	7.76

Tabla i Estadísticas sobre indización de artículos es español en LILACS

Tarea de indización de LILAC e IBECS en BioASQ

El Índice Bibliográfico Español en Ciencias de la Salud (IBECS: http://ibecs.isciii.es) es una base de datos bibliográfica desarrollada por la BNCS en colaboración con Bireme (organismo perteneciente a la Organización Panamericana de la Salud, OPS) utilizando para ello la misma metodología y tesauro que utilizan la base de datos LILACS y Medline. Esto hace que pueda utilizarse un mismo lenguaje de interrogación en todas ellas. IBECS recoge publicaciones de ciencias de la salud españolas que han pasado criterios de calidad en cuanto a normas previamente establecidas. Incluye revistas de diferentes áreas de Ciencias de la Salud, tales como Medicina (incluyendo Salud Pública, Administración Sanitaria y Especialidades), Farmacia, Odontología, Psicología, Enfermería y Fisioterapia.

La base de datos IBECS recoge más de 180 revistas españolas desde el año 2000 en adelante. Con una actualización semanal, actualmente son más de 150.000 registros disponibles de los que más de 148.000 están indizados con el tesauro DeCS. Los no indizados, que forman parte del IBECS Express, irán poco a poco indizándose.

IBECS utiliza la misma metodología de descripción que la base de datos LILACS, desarrollada por BIREME y a su vez compatible con MEDLINE. Esto ha permitido la integración y consulta simultánea de todas ellas a través del portal http://lilacs.bvsalud.org/ , que permite acceder y consultar el referente internacional de publicaciones producidas en países de habla hispana.

Organización de la campaña

Como ya hemos dicho, la campaña que describimos se ocupa de la clasificación a gran escala de documentos biomédicos es español en conceptos de la ontología DeCs. La campaña utilizará el lapso de tiempo entre la primera aparición de un artículo en LILACS e IBECS y su indexación con términos DeCs para preparar conjuntos de pruebas que consisten en artículos no anotados.

Durante el período de competición (tres meses), los conjuntos de pruebas se publicarán regularmente una vez por semana. Esto permitirá a los participantes mejorar sus sistemas teniendo en cuenta los resultados de las evaluaciones parciales que estarán disponibles al tiempo que permite que los participantes se registren en la campaña en cualquier momento.

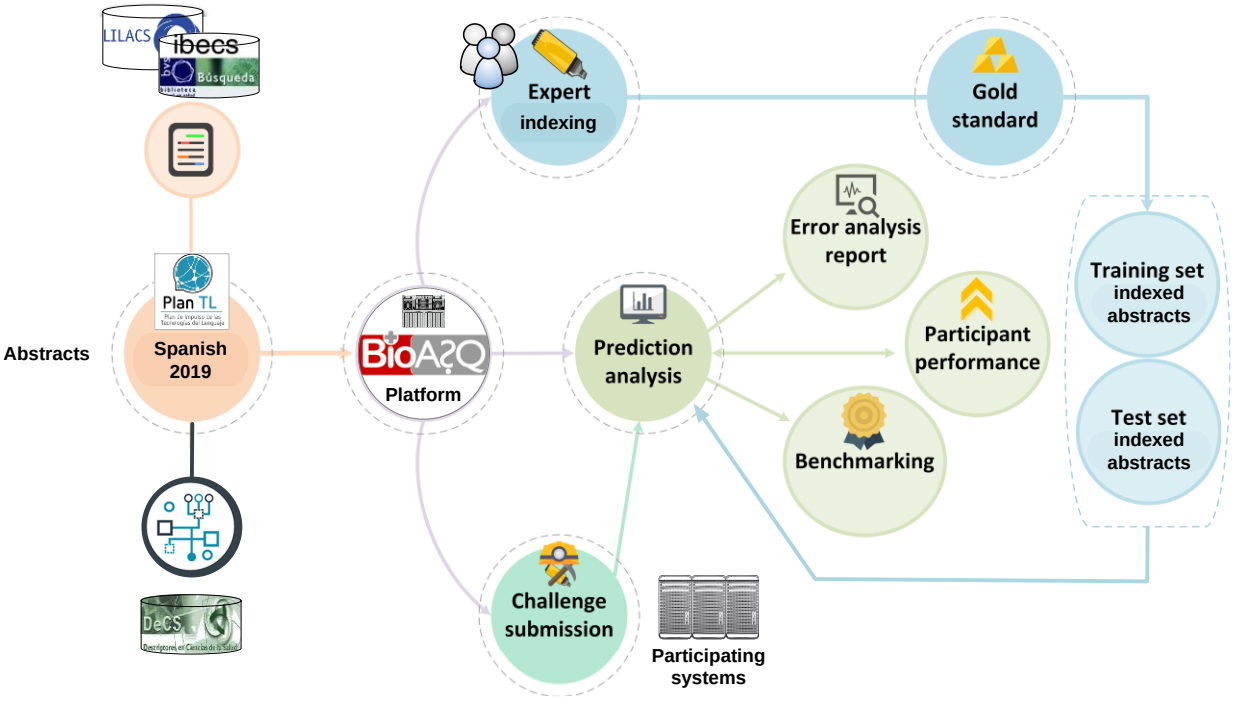

Figura 4 Esquema campaña de indización en BioASQ

Los participantes dispondrán de un tiempo limitado desde el lanzamiento del conjunto de pruebas para enviar sus estimaciones. La evaluación de los sistemas se realiza de forma incremental cada vez que las nuevas anotaciones están disponibles en LILACS e IBECS por curadores humanos. Como datos de entrenamiento, los participantes recibirán todos los artículos indizados previamente con sus respectivas anotaciones DeCs. Para cada artículo en los datos de entrenamiento, se proporciona el título y resumen tal como aparece en LILACS e IBECS y las etiquetas DeCs que se le asignaron. En la fase de prueba (o competición) de la campaña, los datos contienen sólo el título, el resumen, la revista y el año del artículo correspondiente sin más información. Los artículos se proporcionarán en su formato original (texto sin formato).

Para tener una idea de la magnitud de los datos de entrenamiento, la Tabla ii muestra semanalmente, las nuevas incorporaciones a la base de datos IBECS desde noviembre del 2018 a febrero del 2019. La columna 'Ibecs express' contiene las

nueva incorporaciones (y pendientes de indizar) y la columna 'indizados' muestra los artículos indizados esa semana.

Mes	Fecha	Ibecs express	Indizados
Nov-18	02/11 a 08/11	684	71
	12/11 a 16/11	524	136
	19/11 a 26/11	427	148
	26/11 a 30/11	435	274
Dic-18	03/12 a 07/12	329	255
	10/12 a 14/12	446	145
	15/12 a 21/12	437	109
	26/12 a 04/01	177	172
Ene-19	01/1 a 11/01	390	189
	14/01 a 18/01	427	221
	21/01 a 25/01	525	175
	28/01 a 01/02	448	246
Feb-19	04/02 a 08/02	293	258
	11/02 a 15/02	417	320
	18/02 a 22/02	371	360

Tabla ii Nuevas incorporaciones semanales en la base de datos IBECS

Para la evaluación de los sistemas participantes en la tarea, se consideran dos medidas, una plana y una jerárquica. La principal diferencia entre ellos es que esta última toma en cuenta las relaciones en la jerarquía dada, penalizando más las clasificaciones erróneas en ramas distantes de la jerarquía. Ambas medidas son aplicables para la evaluación de todos los tipos de clasificadores. La medida plana que se utiliza es la medida micro-F1, que es una medida basada en etiquetas[17] . La medida jerárquica es la LCaF[18].

Para tener una referencia de partida para las evaluaciones, se suministrará un sistema de referencia (*baseline system* en inglés). El sistema consiste en traducir el resumen al inglés utilizando un sistema de traducción automático entrenado con corpus paralelo inglés/español [8] y utilizar el Medical Text Indexer de la NLM.

Discusión

[17] Tsoumakas G, Katakis I, Vlahavas IP. Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook: 2010. p. 667–85

[18] Kosmopoulos A, Partalas I, Gaussier E, Paliouras G, Androutsopoulos I. Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Mining and Knowledge Discovery. 2014; 29:1–46.

Con respecto a los resultados del reto y al impacto esperado de la campaña BioASQ-es, el principal objetivo a largo plazo es impulsar significativamente la investigación en sistemas y métodos de información que apunten a su vez a un mejor acceso a la información biomédica en español. El impacto potencial de este desarrollo es enorme y afecta a los expertos biomédicos, a las empresas que prestan servicios en este sector, incluidos los proveedores de tecnología de la información y, en última instancia, a todos los que se beneficiarán de la mejora de los procesos biomédicos. En el camino hacia este objetivo, BioASQ-es busca obtener resultados significativos a corto plazo: facilitar una mejor comprensión de las actuales tecnologías de indexación semántica y su aplicación en español; concienciar la comunidad biomédica sobre la posibilidad de una mejora significativa de su trabajo, utilizando sistemas de información inteligentes con especial énfasis a los proveedores de datos; la creación de datos de referencia; y, por último, la integración de la campaña en una infraestructura de prestigio y bien conocida proporciona, sin duda, una excelente base para futuros trabajos de investigación en los campos de la indización semántica biomédica en español.

Es de esperar también que los resultados obtenidos pueden adaptarse para indexar otros documentos en salud, como guías de práctica clínica e incluso para anotar semánticamente páginas web de referencia en el ámbito.

Referencias

[1] Villegas, Marta, et al. "Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje." *Procesamiento del Lenguaje Natural* 59 (2017): 141-144.

[2] Huang, Chung-Chi, and Zhiyong Lu. "Community challenges in biomedical text mining over 10 years: success, failure and the future." *Briefings in bioinformatics* 17.1 (2015): 132-144.

[3] Chapman, Wendy W., et al. "Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions." (2011): 540-543.

[4] Primo-Peña, Elena, and José-Manuel Estrada-Lorenzo. "Las bases de datos bibliográficas españolas, un instrumento para el conocimiento y la difusión de la producción científica." Seminarios de la Fundación Española de Reumatología 10.4 (2009): 132-141.

[5] Mariani J, Paroubek P, Francopoulo G, Hamon O. Rediscovering 15 years of discoveries in language resources and evaluation: The LREC anthology analysis. InProceedings of LREC 2014 May (pp. 26-31).

[6] Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, Almirantis Y. An

overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics. 2015 Dec;16(1):138.

[7] Mork JG, Demner-Fushman D, Schmidt S, Aronson AR. Recent Enhancements to the NLM Medical Text Indexer. InCLEF (Working Notes) 2014 Sep 15 (pp. 1328-1336).

[8] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. Medinfo. 2004 Sep;89

[9] Soares F, Becker K. UFRGS Participation on the WMT Biomedical Translation Shared Task. InProceedings of the Third Conference on Machine Translation: Shared Task Papers 2018 (pp. 662-666).